K-MEANS CLUSTERING AlGORITHM USING INITIALIZATION AND NORMALIZATION METHODS

[1]Rawinder Kaur,[2]Prabhjot Kaur
[1]Student,[2]Assistant Professor
[1]Department of Computer Science Engineering
[1]Patiala Institute Of Engineering & Technology,Patiala,Punjab

***Abstract*—**K means is a very popular algorithm for clustering data in data mining. It is widely used in many research areas because of its efficiency and simplicity. The paper reflects the methods such as normalization and initialization in order to improve the efficiency, accuracy and cluster quality of k- means algorithm.

***Keywords*—clustering; k-means; normalization; intialization**

I. INTRODUCTION

Data clustering is the technique of clustering the data into different groups and these formed groups are known as Clusters in Data mining [1]. Data elements are clustered into different groups based on the similarities and dissimilarities. Basic motive is to keep the similar data elements together in one group. The elements in one cluster have similar properties or similar behavior. For clustering the data we have many clustering algorithms [2]. K-means [3] is one of them that is widely used in research areas because its simplicity, scalability, flexibility and capability to handle large data sets. In k-means[10] we first arbitrary select mean values also called centroid from the given data set. Basically the number of mean values depends on the number of clusters we want to make that is if we want to make two clusters we select two mean values. Then we make clusters or we adjust other data elements in to the clusters by measuring the Euclidean distance[4] of the data element from that mean values. Basically we check the least distance and keep the data element in particular cluster which contain that mean value from where data element is having the least distance. After adjusting all the elements to different clusters we find the mean value of the data elements in the particular cluster in order to find again the mean then adjust elements again according these new mean values. This process continues and we will stop when we found no change in the clusters. In this way we get the final clusters. There were some weaknesses with this traditional k-means algorithm as we arbitrary choose mean values from the data set. So, if next time we choose different mean values from the previous one then it will give different clusters from the previous one. So, with different values it gives different clusters. So, with this we do not know with which mean values it will give accurate results. So, this traditional k-means algorithm doesn't give good quality of clusters. There are many ways and methods to improve the k-means algorithm in order to make it more efficient so that it would give better quality of clusters. This paper aims to provide various methods and ways to improve the k-means method so that it would become more utilizable, more efficient for research purposes in the coming future. Basically in this paper we are going to discuss the ways to normalize the data set before making the clusters and also we will discuss the various initialization techniques that we can use to make better quality of clusters.

II. NORMALIZATION

In data transformation, data is transformed into the form appropriate for mining. Normalization [5] is one of the data transformation strategies. In Normalization [6] data attributes are scaled so as to fall within a particular smaller range such as -1.0 to 1.0, or 0.0 to 1.0. For distance based methods, normalization helps to prevent attributes with initially large ranges. There are many normalization methods such as min-max normalization, z-score normalization, decimal scaling.

A. Min-max normalization

Min-max normalization [7] is the normalization technique that is used to perform the transformation on the original data. Suppose we have minimum and maximum value as min A and max A of an attribute. Min-max normalization [8][11] maps a value v of attribute A to v' as follows:

"Equation1" is v'= (v-min A) / (max A-min A)

Example 1
Suppose that the minimum and maximum values for the attribute income are $11000 and $90000, respectively.

By min-max normalization, a value v of $70000 for income is transformed to v' as follows:

v'= (v-min A)/ (max A - min A)
v'= (70000-11000) / (90000-11000)
v'=59000/79000
v'=0.746

So, it is clear that value has been mapped to the range [0.0,1.0]

B. Z-score normalization

Z-score normalization [8] is the normalization technique that is also used to perform the transformation on the original data. It is also called zero-mean normalization. In this normalization values for an attribute A, are normalized based on the mean and standard deviation. A value v of attribute A is normalized to v' as follows:

"Equation2" is v'= (v-m)/σ,

where m and σ are the mean and standard deviation respectively of an attribute A.

Mean is the average of the values of an attribute calculated as m = (v1+v2+…+vn) /n,

where n is the number of values.

Standard deviation is the square root of the variance of an attribute A calculated as:

$$\sigma^2 = 1/n \sum_{i=1}^{n} (vi - m)^2$$

This method of normalization is useful when the actual minimum and maximum of attribute A are unknown, or when there are outliers that dominate the min-max normalization.

Example 2

Suppose we have the mean and standard deviation of the values for the attribute income $52000 and $16000, respectively. With z-score normalization, a value v $71000 for income is transformed to v' as follows:

v'= (v-m)/σ

v'= (71000-52000)/16000

= 19000/16000

= 1.1875

C. Decimal scaling

Normalization by decimal scaling [8] normalizes the data by moving the decimal point of values of attribute A. The number of decimal points moved depends on the maximum absolute value of attribute A. A value v of attribute A is normalized to v' as follows:

"Equation3" is

$$v' = \frac{v}{10^j}$$

Example 3

Suppose we have values for the attribute A ranging from -976 to 907. The maximum absolute value of the attribute A is 976. So, to normalize the data by decimal scaling, we will divide each value by 1000 (i.e. j=3) so that -976 normalize to -0.976 and 907 normalize to 0.907, which is done as follows:

$$v' = \frac{v}{10^j}$$

For v = 976, v'=976/1000

v'=0.976

Similarly for v=907, v'=907/1000

v'=0.907

III. INITIALIZATION METHODS

In k-means initialization [9] means the step that we used to choose the centroid for the cluster. In traditional k-means we arbitrary choose the centroid, which doesn't give the good quality clusters. So, there are various initialization methods for better initialization of the cluster.

A. Method 1

Step1. First calculate the average score of each data set as follows:

$$\text{Average score}(x_i) = \frac{\sum_1^m x_i}{m}$$

Step2. Now sort the data based on this average score.

Step3. Divide the sorted data points into equal k groups as follows

Step3.1 Group size = n/k

Step3.2 Group l contains data points as

$X_{(l-1)*groupsize+1}$ to $X_{l*groupsize}$,
$for\ 1 \le l \le k-1$

Step3.3 Keep the remaining data points in kth group i.e. the next group.

Step4. Now calculate the mean value of each group to get the initial centroid for that group.

B. Method 2

Step1. First calculate the average score of each data set as follows:

$$\text{Average score}(x_i) = \frac{\sum_1^m x_i}{m}$$

Step2. Now sort the data based on this average score.

Step3. Divide the sorted data points into equal k groups as follows:

Step3.1 Group size = n/k

Step3.2 Group l contains data points as

$$X_{(l-1)*groupsize+1} \text{ to } X_{l*groupsize} ,\quad for\ 1 \le l \le k-1$$

Step3.3 Keep the remaining data points in kth group i.e. the next group

Step4. Now calculate the median value of each group to get the initial centroid for that group.

So, in this method we calculate the median in order to get the centroid of the group. Median is the middle value but if we have even number of values for an attribute then we will calculate the average of the middle two values in order to get the median.

C. Method 3

Step1. First calculate the average score of each data set as follows:

$$\text{Average score}(x_i) = \frac{\sum_1^m x_i}{m}$$

Step2. Now sort the data based on this average score.

Step3. Divide the sorted data points into equal k groups as follows

Step3.1 Group size = n/k

Step3.2 Group l contains data points as

$$X_{(l-1)*groupsize+1} \text{ to } X_{l*groupsize} ,\quad for\ 1 \le l \le k-1$$

Step3.3 Keep the remaining data points in kth group i.e. the next group.

Step4. Now calculate the mode value of each group to get the initial centroid for that group.

Example 4

Using initialization method 1

Suppose we have following data:

DATA	X1	X2	X3	X4
D1	1	5	4	3
D2	6	7	5	8
D3	9	1	5	4
D4	6	2	9	5
D5	8	9	11	6
D6	25	5	4	3
D7	2	10	8	22
D8	4	9	10	8

Step1. Calculating average score

Average Score for Data set D1= (1+5+4+3)/4=13/4=3.25

Average Score for Data set D2= (6+7+5+8)/4=26/4=6.5

Average Score for Data set D3= (9+1+5+4)/4=19/4=4.75

Average Score for Data set D4= (6+2+9+5)/4=22/4=5.5

Average Score for Data set D5= (8+9+11+6)/4=34/4=8.5

Average Score for Data set D6= (25+5+4+3)/4=37/4=9.25

Average Score for Data set D7= (2+10+8+22)/4=42/4=10.5
Average Score for Data set D8= (4+9+10+8)/4=31/4=7.75

Step2. Sorting of the data set values
Sorted List is as follows:

X1	X2	X3	X4	X5	X6	X7	X8
3.25	4.75	5.5	6.5	7.75	8.5	9.25	10.5

Step3. Suppose number of groups would be k=2
Group size = n/k=8/2=4, n is the total number of elements
It means each group will have four elements
Group l=1, elements or values that it will contain would be:

x_1 to x_4

And rest of the elements will move to the second group
Group 1 will have following elements:

X1	X2	X3	X4
3.25	4.75	5.5	6.5

Group 2 will have following elements:

X5	X6	X7	X8
7.75	8.5	9.25	10.5

Step4. Calculating mean value for initializing the centroid as follows:
For Group 1
Mean= (3.25+4.75+5.5+6.5)/4=20/4=5
For Group 2
Mean= (7.75+8.5+9.25+10.5)=36/4=9

So, 5 is the centroid for the group 1 and 9 is the centroid for the group 2
So, these centroids are calculated according to the mean of the values that are contained in to the group

Example 5
Using initialization method 2
Taking the same data:

DATA	X1	X2	X3	X4
D1	1	5	4	3
D2	6	7	5	8
D3	9	1	5	4
D4	6	2	9	5
D5	8	9	11	6
D6	25	5	4	3
D7	2	10	8	22
D8	4	9	10	8

Step1. Calculating average score
Average Score for Data set D1= (1+5+4+3)/4=13/4=3.25
Average Score for Data set D2= (6+7+5+8)/4=26/4=6.5
Average Score for Data set D3= (9+1+5+4)/4=19/4=4.75
Average Score for Data set D4= (6+2+9+5)/4=22/4=5.5
Average Score for Data set D5= (8+9+11+6)/4=34/4=8.5
Average Score for Data set D6= (25+5+4+3)/4=37/4=9.25
Average Score for Data set D7= (2+10+8+22)/4=42/4=10.5
Average Score for Data set D8= (4+9+10+8)/4=31/4=7.75

Step2. Sorting of the data set values
Sorted List is as follows:

X1	X2	X3	X4	X5	X6	X7	X8
3.25	4.75	5.5	6.5	7.75	8.5	9.25	10.5

Step3. Suppose number of groups would be k=2
Group size = n/k=8/2=4, n is the total number of elements
It means each group will have four elements
Group l=1, elements or values that it will contain would be:

x_1 to x_4

And rest of the elements will move to the second group
Group 1 will have following elements:

X1	X2	X3	X4
3.25	4.75	5.5	6.5

Group 2 will have following elements:

X5	X6	X7	X8
7.75	8.5	9.25	10.5

Step4. Calculating median value for initializing the centroid as follows:
Median is the middle value of the list. For even numbers of list middle value will be calculated as the average of the two middle values

For Group 1
Median = (4.75+5.5)/2 = 10.25/2 = 5.125
For Group 2
Median = (8.5+9.25)/2 = 17.75/2 = 8.875

So, 5.125 is the centroid for the group 1 and 8.875 is the centroid for the group 2

Example 6
Using initialization method 3
Mode is used to calculate the centroid incase we have the repeating values in the group. Then we take that repeating value as the centroid of that group.

For example we have two groups as follows:
Group 1

X1	X2	X3	X4
3.25	4.6	4.6	5.5

In this we have values at X2 and X3 are repeating values so,
So, mode = 4.6, this would be the centroid of the Group 1

Similarly For Group 2

X5	X6	X7	X8
6.5	6.5	8.5	9.25

So, now in this group centroid would be 6.5 because it is the repeating value in the group. Keep in mind that if there is no repeating value then there can be no centroid by this method.

IV. CONCLUSION

K-means is one of the most widely used algorithm in data mining because of its simplicity and efficiency. As in the traditional k-means algorithm there were many limitations that don't give the good quality clusters. In traditional k-means we arbitrary choose the centroids for the group. So, if for next time we choose different centroid the final clusters would contain different elements. So, in this way it depends upon the initialization that what values we initially choose according to those only final clusters would form. So, in this paper we present the various initialization methods that we can take to choose the centroid of the group. This will generate the good quality clusters. In addition to this we had also discussed the various normalization techniques such as min-max normalization, z-score normalization, decimal scaling. It is very important to normalize our data of different ranges before making clusters. Normalization will normalize the data in to a particular range. This will also be very helpful for generating good quality clusters.

REFERENCES

[1] Sukhvir Kaur, "Survey of different data clustering algorithms", *International Journal of Computer Science and Mobile Computing*, Vol.5 Issue.5, May- 2016, pg. 584-588.

[2] Arpita Nagpal, Aman Jatain, Deepti Gaur,"Review based on Data Clustering Algorithms", *Proceedings of 2013 IEEE Conference on Information and Communication Technologies*, (ICT 2013).

[3] Shruti Kapil, Meenu Chawla, "Performance Evaluation of K-means Clustering Algorithm with Various Distance Metrics", *1st IEEE International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES)*, 2016.

[4] Unnati R. Raval, Chaita Jani, "Implementing & Improvisation of K-means Clustering Algorithm", *International Journal of Computer Science and Mobile Computing*, Vol.5 Issue.5, May- 2016, pg. 191-203.

[5] Shruti Gupta, Abha Thakral, Shilpi Sharma , "Novel Technique for Prediction Analysis Using Normalization for an improvement in K-means Clustering", *IEEE International Conference of Information Technology (InCITe) – The Next Generation IT Summit*, 2016.

[6] Akanksha Choudhary, Mr. Prashant Sharma,"Mr. Manoj Singh Improving K-Means Through Better Initialization And Normalization", *2016 Intl. Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Sept. 21-24, 2016, Jaipur, India.

[7] Navdeep Kaur, Krishan Kumar, "Normalization Based K-means Data Analysis Algorithm", *International Journal of Advanced Research in Computer Science and Software Engineering*, Volume 6, Issue 6, June 2016

[8] Ismail Bin Mohamad , Dauda Usman , "Standardization and Its Effects on *K*-Means Clustering Algorithm", *Research Journal of Applied Sciences, Engineering and Technology 6(17): 3299-3303*, 2013.

[9] Wei Du, Hu Lin, Jianwei Sun, Bo Yu and Haibo Yang, "A New Projection-based K-Means Initialization Algorithm", *Proceedings of 2016 IEEE Chinese Guidance, Navigation and Control Conference*, August 12-14, 2016 Nanjing, China.

[10] Malwinder Singh, Meenakshi Bansal, "A Survey on Various K-Means algorithms for Clustering", *IJCSNS International Journal of Computer Science and Network Security*, VOL.15 No.6, June 2015.

[11] Rishikesh Suryawanshi , Shubha Puthran , "A Novel Approach for Data Clustering using Improved K-means Algorithm", *International Journal of Computer Applications (0975 – 8887)*, Volume 142 – No.12, May 2016

[12] D. Virmani, S. Taneja, G. Malhotra, "Normalization based K means Clustering Algorithm*", International Journal of Advanced Engineering Research and Science*, Vol-2, Issue-2,. 2015